WORKING PAPER
ALFRED P. SLOAN SCHOOL OF MANAGEMENT

REDUCING THE VARIANCE OF SOJOURN TIMES IN MULTICLASS QUEUEING SYSTEMS

Lawrence M. Wein

MIT Working Paper # 3093-89MS
November 1989

MASSACHUSETTS
INSTITUTE OF TECHNOLOGY
50 MEMORIAL DRIVE
CAMBRIDGE, MASSACHUSETTS 02139

REDUCING THE VARIANCE OF SOJOURN TIMES IN MULTICLASS QUEUEING SYSTEMS

Lawrence M. Wein

MIT Working Paper # 3093-89MS
November 1989

REDUCING THE VARIANCE OF SOJOURN TIMES IN MULTICLASS QUEUEING SYSTEMS

Lawrence M. Wein

Sloan School of Management, M.I.T.

We consider the bi-criteria scheduling problem of minimizing a convex combination of the mean and variance of sojourn times in three queueing systems (a single queue, a two station closed network, and a two station network with controllable inputs) populated by various customer types. The sum of the squares of the pairwise differences in mean sojourn times among the various customer types is used as a surrogate for the sojourn time variance. Brownian approximations to these three scheduling problems are solved, and the solutions are interpreted in order to obtain scheduling policies. Simulation results show that in the network settings, in contrast to the single queue case, there are priority sequencing policies that significantly reduce the variance of sojourn times relative to the first-come first-served policy.

Subject classification: Production/scheduling: priority sequencing in a stochastic job shop. Queues: Brownian models of network scheduling problems.

November 1989

REDUCING THE VARIANCE OF SOJOURN TIMES IN MULTICLASS QUEUEING SYSTEMS

Lawrence M. Wein

Sloan School of Management, M.I.T.

The objective in scheduling most multiclass queueing systems is to minimize the mean sojourn time of customers. However, in many practical situations, it is also desirable to keep the variance of the sojourn times as small as possible. For example, in many manufacturing systems, all arriving jobs are quoted the same due-date lead time (due-date minus arrival date), and so a smaller sojourn time variance will lead to better due-date performance. In some service systems, the desire to treat all customers as equitably as possible dictates that a small sojourn time variance is required. In many production facilities, low variability in sojourn times leads to easy coordination with its downstream customers and upstream suppliers.

For single server, single class queueing systems, the first-come first-served (FCFS) policy minimizes sojourn time variance over the class of work-conserving sequencing policies (see Kingman [15]). Furthermore, Groenevelt [4] has shown that FCFS is effective in reducing sojourn time variance in single server, multiclass queues. It is also well known that the shortest expected processing time (SEPT) rule (see Klimov [18], for example) minimizes the mean sojourn time in many single server, multiclass queueing systems. The SEPT rule, however, can lead to a very large sojourn time variance; when the load on the system is very heavy, customers with the largest expected processing times may sit in the queue for a long period of time, while customers with the shortest expected processing time will usually move through the system relatively quickly. Thus, a fundamental tradeoff exists in single server, multiclass queues between system *efficiency* (minimizing mean sojourn time) and system *equitability* (minimizing sojourn time variance). As a result, some authors

(see Jackson [10]-[12], Kleinrock [16], and Shanthikumar [22], for example) have suggested and analyzed alternative sequencing policies that attempt to address this tradeoff in an effective manner.

However, very little is known about this tradeoff in a network setting. Indeed, just obtaining sojourn time variances in single class queueing networks under the FCFS policy is an arduous task (see Chapter 4 of Walrand [24] and references therein). In this paper, the tradeoff is assessed by analyzing a bi-criteria scheduling problem for three multiclass queueing systems. Before describing the various criteria, we first need to define the basic queueing network underlying all three queueing systems. Using the terminology of Kelly [14], we assume the queueing network is visited by various *types* of customers, each with its own arbitrary deterministic route through the system. In a manufacturing setting, a customer type represents a particular product that is produced by the facility. A different customer *class* will be defined for each combination of type and stage of completion, and each customer class is served at a particular single server station in the network, and has its own general service time distribution.

The first of the three queueing systems is a single server, multiclass queue, where each of the various types of customers has its own independent renewal input process. The second system is a closed network, where the total population in the network is held constant, and when a customer departs the network, the type of the new customer is chosen exogenously (in a Markovian or deterministic fashion) according to a specified product mix. In both of these systems, the scheduling decision is to dynamically decide which customer class to serve next at each station; we will refer to these decisions as *sequencing* decisions. The third system is a network with controllable inputs, where the scheduler also decides when to release the next customer into the network. It is assumed that there is a specified expected average throughput rate that must be maintained, and there is an infinite line of customers lined up outside the network waiting to gain entrance. The type of each entering customer is again exogenously specified according to a particular product mix.

Since analyzing the variance of sojourn times is so difficult in a multiclass queueing network, we propose a surrogate measure for the bi-criteria scheduling problems. In particular, define the *surrogate variance* of the sojourn times as the sum of the squares of the pairwise differences in mean sojourn times among the various customer types. Thus, if every customer type has the same mean sojourn time, then there is zero surrogate variance. The first of the two criteria in each of the three minimization problems is the surrogate variance. In the single station queue and the network with controllable inputs, the second criterion is the mean number of customers in the system, which is proportional to the mean sojourn time in both systems by Little's formula (see Little [19]). In the closed network, the second criterion is the mean idleness rate at an arbitrarily chosen server, which is inversely proportional to the mean throughput rate of the network. Hence minimizing the mean idleness rate of a server will minimize the mean sojourn time of customers. In all three problems, a weight c will be put on the second criterion, and the objective is to minimize the weighted average of the two criteria. When $c = 0$, the objective is to minimize the surrogate variance, and when $c \to \infty$, the objective is to minimize the mean sojourn time.

If the primary objective of the queueing system is to equalize the sojourn times across the various customer types, then minimizing the surrogate variance is the most appropriate criterion. Moreover, a reduction in the surrogate variance should lead to a reduction in the actual sojourn time variance, unless the reduction in surrogate variance results in a large increase in the mean sojourn times. In such cases, solving the bi-criteria scheduling problem for various values of c should lead to a reduction in sojourn time variance.

The three bi-criteria scheduling problems will be analyzed using the Brownian network model introduced by Harrison [7], which approximates a multiclass queueing network with dynamic scheduling capability under balanced heavy loading conditions. This model allows the bi-criteria scheduling problems to be approximated by control problems for Brownian motion. When $c \to \infty$, the three scheduling problems described above have been analyzed

in Harrison [7] (single queue), Harrison and Wein [9] (closed network), and Wein [25]-[26] (network with controllable inputs) by solving and interpreting a reformulation of the approximating Brownian control problem.

In this paper, we generalize these analyses to include the bi-criteria objective for any value of c. For the single queue, we find in Section 1 that FCFS is effective in minimizing the surrogate variance, which is not surprising in light of Groenevelt's [4] results. The closed network and network with controllable inputs are analyzed in Sections 2 and 3, respectively, and a simulation experiment involving a two-station network is undertaken in Section 4 that demonstrates the effectiveness of the procedure.

The simulation results imply that, as in the single server case, the tradeoff between system efficiency and system equitability exists in the network setting. However, whereas FCFS minimizes sojourn time variance in the single server case, we find that it is easily outperformed on this dimension in the network case. This should not be surprising, since some customer types may have longer routes (and/or processing times) than other customer types, and FCFS will generally cause the types with the longer routes to have longer sojourn times.

The simulation results suggest that reducing the surrogate variance does indeed seem to be an effective surrogate for reducing sojourn time variance. Also, the nature of the tradeoff between the mean sojourn time and sojourn time variance is problem specific, and depends on the magnitude of the surrogate variance under the policy that minimizes the mean sojourn time. If the surrogate variance is small, then a further large reduction in sojourn time variance is not possible. However, if the surrogate variance is large, then a further large reduction in sojourn time variance may be possible. We simulated our test example under two sets of data in order to illustrate these two cases. It is interesting to note that in both cases, a sequencing policy was found that simultaneously reduced the mean sojourn time and the sojourn time variance with respect to FCFS.

In the simulation study, the closed network and network with controllable inputs

were tested against an open network that used deterministic input. Not surprisingly, the reduction in sojourn time variance from the open network to the other two networks (using the FCFS sequencing policy in all cases) easily dominated the reduction in sojourn time variance among the various sequencing policies (while holding the release policy fixed).

In summary, the easiest and most commonly used priority sequencing policy (FCFS) maximizes system equitability in single server queues, and therefore, if one wants to maximize system efficiency with respect to FCFS, then one sacrifices some equitability. Our results suggest that in a network setting, there exists sequencing policies that improve both system efficiency and system equitability with respect to FCFS.

1. A Single Server Queue

Consider a single server queue visited by K customer classes who arrive according to independent renewal processes with average arrival rates $\lambda_k, k = 1, ..., K$. Each customer class has a general service time distribution with mean m_k and finite variance, and each customer exits the system after receiving service. Therefore, the traffic intensity of the queueing system is $\rho = \sum_{k=1}^{K} \lambda_k m_k$. Let $Q_k(t)$ be the number of class k customers in the system at time t, and let $I(t)$ be the cumulative amount of time that the server is idle up to time t. The Brownian model assumes the existence of a large integer n such that $\sqrt{n}(1-\rho)$ is positive and of moderate size: a representative example is $\rho = .9$ and $n = 100$. The system parameter n is used to define the rescaled processes $Z = (Z_k)$ and U by $Z_k(t) = Q_k(nt)/\sqrt{n}$ and $U(t) = I(nt)/\sqrt{n}$ for $t \geq 0$, and the Brownian control problem is obtained by letting the parameter n approach infinity.

Since the proposed scheduling policy depends only on the solution to the workload formulation, we will go directly to the workload formulation of the Brownian control problem that approximates this scheduling problem. In Section 6 of Harrison [x], it is shown that the problem (the objective function will be stated shortly) is to choose RCLL (right

continuous with left limits) processes $Z = (Z_k)$ and U such that

$$\sum_{k=1}^{K} m_k Z_k(t) = B(t) + U(t) \quad \text{for all} \quad t \geq 0. \tag{1.1}$$

$$Z(t) \geq 0 \quad \text{for all } t \geq 0, \quad \text{and} \tag{1.2}$$

$$U \text{ is nondecreasing with } U(0) = 0, \tag{1.3}$$

where B is a one-dimensional Brownian motion process with a particular drift and covariance, and with $B(0) = 0$. For all the problems addressed in this paper, the form of the scheduling policy is independent of the parameters of the various Brownian motion processes. These parameters are complicated expressions involving the first and second moments of the interarrival times and service times, the product mix, and the routing information; we refer readers to Harrison [7] for their derivation. Also, it will suffice to assume that Z and U are nonanticipating with respect to the Brownian motion B. Similar assumptions are made in later sections, and readers are referred to the earlier papers in this area for a justification.

By Little's formula, our objective function can be written as

$$\limsup_{T\to\infty} \frac{1}{T} E\left[\int_0^T \left(c\sum_{k=1}^{K} Z_k(t) + \frac{1}{2}\sum_{j=1}^{K}\sum_{k=1}^{K} \left(\frac{Z_j(t)}{\lambda_j} - \frac{Z_k(t)}{\lambda_k}\right)^2 \right) dt\right]. \tag{1.4}$$

The first term in (1.4) is proportional to the mean steady-state sojourn time, and the second term is the surrogate sojourn time variance. By the argument in Section 6 of Harrison [7], the solution to this problem is to choose

$$U(t) = -\inf_{0\leq s\leq t} B(s) \tag{1.5}$$

and to find, for each time t, the solution $Z^*(t)$ to the quadratic program

$$\min \quad c\sum_{k=1}^{K} Z_k(t) + \frac{1}{2}\sum_{j=1}^{K}\sum_{k=1}^{K} \left(\frac{Z_j(t)}{\lambda_j} - \frac{Z_k(t)}{\lambda_k}\right)^2 \tag{1.6}$$

$$\text{subject to} \quad \sum_{k=1}^{K} m_k Z_k(t) = W(t), \tag{1.7}$$

$$Z(t) \geq 0, \tag{1.8}$$

where the one-dimensional *workload process* $W(t)$ is defined by

$$W(t) = B(t) - \inf_{0 \le s \le t} B(s), \tag{1.9}$$

and represents the (scaled) total amount of work in the system. Furthermore, such a solution is referred to as a *pathwise solution* to the workload formulation (1.1)-(1.4), in that it minimizes the objective for all times t with probability one.

Proposition 1. *When $c = 0$, the solution to (1.6)-(1.8) is*

$$Z_k^*(t) = \frac{\lambda_k W(t)}{\rho} \text{ for } k = 1, ..., K \text{ and } t \ge 0. \tag{1.10}$$

Proof. Since the objective function is convex for any value of $c \ge 0$ and the constraints are linear, it follows that the first-order Karush-Kuhn-Tucker conditions are sufficient conditions for optimality (see Avriel [1], for example). These conditions are to find Z_k^* and π^* that satisfy (1.7)-(1.8) and

$$c + \frac{2(K-1)Z_k}{\lambda_k^2} - \frac{2}{\lambda_k} \sum_{j \ne k} \frac{Z_j}{\lambda_j} - \pi m_k \ge 0 \text{ for } k = 1, ..., K, \tag{1.11}$$

$$\left[c + \frac{2(K-1)Z_k}{\lambda_k^2} - \frac{2}{\lambda_k} \sum_{j \ne k} \frac{Z_j}{\lambda_j} - \pi m_k \right] Z_k = 0 \text{ for } k = 1, ..., K. \tag{1.12}$$

Readers may verify that when $c = 0$, a solution to (1.7)-(1.8), (1.11)-(1.12) is $\pi^* = 0$ and (1.10). ∎

Thus, in the Brownian limit, the queue lengths of the various classes are in direct proportion to their respective arrival rates when $c = 0$. Johnson [13] and Peterson [20] have shown that this same proportionality also holds under the FCFS policy in the heavy traffic limit. Thus, our analysis suggests that the FCFS policy is effective in reducing sojourn time variance in single server, multiclass queues. It is encouraging to note that our analysis concurs with a related result using exact methods (see Groenevelt [4]).

In the limiting case when $c \to \infty$, Harrison [7] shows that the proposed sequencing policy is the shortest expected processing time rule, which again concurs with exact results

(see, for example Klimov [18]). When $0 < c < \infty$, we propose the following procedure to obtain an effective sequencing policy. First, find the solution $(Z_k^*(t), \pi^*)$ to (1.7)-(1.8), (1.11)-(1.12). This solution Z^* will yield a unique value of the unscaled workload vector in steady-state by the conversation law in Kleinrock [17], and then the unique parameters for Kleinrock's delay dependent discipline (see Kleinrock [16]-[17]) can be derived that will achieve this workload vector. In particular, the synthesis algorithm of Wood and Sargeant [30] offers a simple iterative algorithm (where a quadratic equation is solved at each step) to derive the necessary parameters (see also Federgruen and Groenevelt [3] for a generalization of this algorithm).

2 A Closed Network

For the networks in this section and the next section, we will index single server stations by $i = 1, 2$, customer classes by $k = 1, ..., K$, and customer types by $j = 1, ..., J$. Since a different customer class is defined for each combination of customer type and stage of completion, we have $K \geq J$. Let $\tau(j)$ be the set of classes that correspond to a particular stage on the route of type j. The total population size of the closed network considered in this section is always maintained at N, and this is achieved by releasing a new customer into the network whenever one exits. The new customer will be of class k with probability q_k, independent of all previous history, where $\sum_{k=1}^{K} q_k = 1$; alternatively, entering customers can be chosen according to the mix q_k in a deterministic fashion, and the only change in the analysis is the covariance matrix of the underlying Brownian motion. Of course, $q_k > 0$ only for those classes that correspond to the first stage of some customer type's route. We will also let $\bar{q}_j = \sum_{k \in \tau(j)} q_k$, so that entering customers are of type j with probability $\bar{q}_j$.

Recall that each customer type has its own arbitrary deterministic route through the network. In the workload formulation of the Brownian control problem, the routing

information is captured in the quantity M_{ik}, which is the expected remaining processing time at station i for a class k customer until that customer exits the network. For $i = 1, ..., I$, define $v_i = \sum_{k=1}^{K} M_{ik} q_k$, so that v_i is the expected total time over the long run that server i devotes to each newly arriving customer. Define the relative traffic intensity ρ_i for station i to be $v_i / \max\{v_1, v_2\}$ for $i = 1, 2$. Then the balanced heavy loading conditions for the closed network assume the existence of a sufficiently large integer N such that the total population size is N and $N|\rho_1 - \rho_2|$ is of moderate size; a representative example is $N = 10$ and $\rho_1 = \rho_2 = 1$.

As in the previous section, the queue length process and server idleness process need to be rescaled, this time by the population parameter N. Let $Z_k(t) = Q_k(N^2 t)/N$ for $k = 1, ..., K$ and $t \geq 0$, and $U_i(t) = I(N^2 t)/N$ for $i = 1, 2$ and $t \geq 0$. Let $\hat{B}$ be a one-dimensional Brownian motion process with drift μ and variance σ^2, and let

$$\hat{M}_k = \rho_2 M_{1k} - \rho_1 M_{2k} \text{ for } k = 1, ..., K. \tag{2.1}$$

By Propositions 2 and 7 in Harrison and Wein [9], the workload formulation can be expressed as choosing RCLL processes (U_1, U_2, Z_k) that are nonanticipating with respect to $\hat{B}$ such that

$$\sum_{k=1}^{K} \hat{M}_k Z_k(t) = \hat{B}(t) + \rho_2 U_1(t) - \rho_1 U_2(t) \quad \text{for all } t \geq 0, \tag{2.2}$$

$$\sum_{k=1}^{K} Z_k(t) = 1 \quad \text{for all } t \geq 0, \tag{2.3}$$

$$Z_k(t) \geq 0 \text{ for } k = 1, ..., K \text{ and for all } t \geq 0. \text{ and} \tag{2.4}$$

$$U_1 \text{ and } U_2 \text{ are nondecreasing with } U_1(0) = U_2(0) = 0. \tag{2.5}$$

If the number of customers exiting the network per unit of time is λ over the long run and $Z_k(\infty)$ denotes the scaled vector queue length process in steady-state, then by Little's formula, the mean steady-state sojourn time of type j customers is $E[\sum_{k \in r(j)} Z_k(\infty)]/\bar{q}_j \lambda$. Even though the throughput rate of the network is unknown, we can express our bi-criteria

objective as

$$\min \limsup_{T\to\infty} \frac{1}{T} E\left[cU_1(T) + \frac{1}{2}\int_0^T \sum_{i=1}^{J}\sum_{j=1}^{J}\left(\frac{\sum_{k\in\tau(i)} Z_k(t)}{\bar{q}_i} - \frac{\sum_{k\in\tau(j)} Z_k(t)}{\bar{q}_j}\right)^2 dt\right]. \quad (2.6)$$

In the limiting case $c \to \infty$, the solution to this problem (see Harrison and Wein [9] for details) is to choose

$$U_1^*(t) = \frac{1}{\rho_2} \sup_{0\le s\le t} [a^* - \hat{B}(s) + \rho_1 U_2^*(s)]^+, \quad \text{and} \quad (2.7)$$

$$U_2^*(t) = \frac{1}{\rho_1} \sup_{0\le s\le t} [\hat{B}(s) + \rho_2 U_1^*(s) - b^*]^+, \quad (2.8)$$

so that $\hat{B}(t)+\rho_2 U_1(t)-\rho_1 U_2(t), t \ge 0$, is a one-dimensional reflected (or regulated) Brownian motion (abbreviated by RBM; see Harrison [6] for a complete treatment) on the interval $[a^*, b^*]$. The interval endpoints are defined by $a^* = \min_{1\le k\le K} \hat{M}_k$ and $b^* = \max_{1\le k\le K} \hat{M}_k$, and $a^* < 0 < b^*$. Defining

$$\hat{W}^*(t) = \hat{B}(t) + \rho_2 U_1^*(t) - \rho_1 U_2^*(t), \quad (2.9)$$

and assuming (without loss of generality) that the classes $k = 1, \ldots, K$ are ordered so that $\max_{1\le k\le K} \hat{M}_k = \hat{M}_1$ and $\min_{1\le k\le K} \hat{M}_k = \hat{M}_2$, then

$$Z_k^*(t) = \begin{cases} \gamma(t) & \text{if k = 1;} \\ 1-\gamma(t) & \text{if k = 2;} \\ 0 & \text{if k = 3,....K.} \end{cases} \quad (2.10)$$

for all $t \ge 0$, where

$$\gamma(t) = \frac{\hat{W}^*(t) - a^*}{b^* - a^*} \quad \text{for all } t \ge 0. \quad (2.11)$$

As in the single queue case, (Z^*, U^*) is a pathwise solution to problem (2.2)-(2.6). The resulting sequencing policy (see Harrison and Wein [9] for an interpretation of the solution (Z^*, U^*)) is to rank each customer class $k = 1, \ldots, K$ by the index $\hat{M}_k$, and to award higher priority at station 1 (respectively, station 2) to the classes with smaller (respectively, larger) values of this index. Notice that this policy awards bottom priority at each station to the

class that has a positive scaled queue length, thus maintaining consistency with existing heavy traffic limit theorems: see, for example, Whitt [29], Harrison [5], and Reiman [21].

Now let us analyze the other limiting case, when $c = 0$. It is clear that the set of values $\hat{W}$ such that there exist $Z_1, ..., Z_K$ satisfying

$$\hat{W} = \sum_{k=1}^{K} \hat{M}_k Z_k, \tag{2.12}$$

$$\sum_{k=1}^{K} Z_k = 1, \tag{2.13}$$

$$Z_k \geq 0, \text{for } k = 1, ..., K \text{ and} \tag{2.14}$$

$$\frac{\sum_{k \in \tau(i)} Z_k}{\bar{q}_i} - \frac{\sum_{k \in \tau(j)} Z_k}{\bar{q}_j} = 0 \text{ for } i \neq j, \tag{2.15}$$

is a closed interval on the real line that is a subset of the interval $[a^*, b^*]$. Let us denote this interval by $[a^0, b^0]$, where the superscript reminds us that we are analyzing the case $c = 0$. Therefore, any solution (Z^0, U^0) that keeps $\hat{W}(t)$ within the interval $[a^0, b^0]$ will be a pathwise optimal solution to (2.2)-(2.6). There may be many pathwise solutions, and we propose to choose

$$U_1^0(t) = \frac{1}{\rho_2} \sup_{0 \leq s \leq t} [a^0 - \hat{B}(s) + \rho_1 U_2^0(s)]^+, \quad \text{and} \tag{2.16}$$

$$U_2^0(t) = \frac{1}{\rho_1} \sup_{0 \leq s \leq t} [\hat{B}(s) + \rho_2 U_1^0(s) - b^0]^+, \tag{2.17}$$

so that $\hat{W}(t)$ is a one-dimensional RBM on $[a^0, b^0]$. Among the pathwise solutions to (2.2)-(2.6), these values will maximize the mean throughput of the network (see Proposition 2 below), and hence will minimize the mean sojourn time.

In order to find an effective sequencing policy, we find solutions Z^{a^0} and Z^{b^0} that satisfy (2.12)-(2.15) under the two extreme points $\hat{W} = a^0$ and $\hat{W} = b^0$, respectively. All classes that have $Z_k^{a^0} = Z_k^{b^0} = 0$ will be in the higher priority bracket at their respective stations, and all other classes will be in the lower priority bracket. At each station, all classes within the higher priority bracket are given priority over all classes in the lower

priority bracket. Within the higher priority bracket, classes are ranked in the same manner as when $c \to \infty$; this ranking maintains the spirit of minimizing mean sojourn time. A simple dynamic policy (that will be illustrated in the example in Section 4) is employed to attempt to maintain the actual queue lengths of the lower bracket classes in the same relative quantities as Z^{a^0} and Z^{b^0}.

Now let us consider the general bi-criteria case where $0 < c < \infty$. Notice that our solution (U_1, U_2) for both limiting cases (see (2.6)-(2.7) and (2.16)-(2.17)) are of a very special form (they are referred to as *control limit policies* in Harrison and Taksar [8]) and imply that $\hat{W}(t)$ in (2.9) is a RBM on some closed interval. We will restrict ourselves to this class of policies (this restriction will be justified shortly), so that a policy (U_1, U_2) will be characterized by a particular closed interval $[a, b]$. It is clear that for $0 < c < \infty$, the optimal interval endpoints a and b will satisfy $a^* \le a \le a^0$ and $b^0 \le b \le b^*$, where $[a^*, b^*]$ characterizes the solution when $c \to \infty$, and $[a^0, b^0]$ characterizes the solution when $c = 0$.

For a control limit policy (U_1, U_2) characterized by the interval $[a, b]$, the following two propositions are well-known results (see Chapter 5 of Harrison [6]).

Proposition 2.

$$\lim_{T\to\infty} \frac{1}{T} E[U_1(T)] = \begin{cases} \frac{\mu}{\rho_2(e^{\nu(b-a)}-1)} & \text{if } \rho_1 \neq \rho_2; \\ \frac{\sigma^2}{2(b-a)} & \text{if } \rho_1 = \rho_2, \end{cases} \tag{2.18}$$

and

$$\lim_{T\to\infty} \frac{1}{T} E_x[U_2(T)] = \begin{cases} \frac{\mu}{\rho_1(1-e^{-\nu(b-a)})} & \text{if } \rho_1 \neq \rho_2; \\ \frac{\sigma^2}{2(b-a)} & \text{if } \rho_1 = \rho_2, \end{cases} \tag{2.19}$$

where $\nu = 2\mu/\sigma^2$.

Proposition 3. *If* $\hat{W}(t)$ *is a RBM on* $[a, b]$, *then* $\hat{W}$ *has a uniform steady-state distribution on* $[a, b]$ *if* $\rho_1 = \rho_2$, *and otherwise has a truncated exponential steady-state distribution with density function*

$$p(x) = \frac{\nu e^{\nu(x-a)}}{e^{\nu(b-a)} - 1} \quad \text{for } a \le x \le b. \tag{2.20}$$

These results allow us to analyze (2.2)-(2.6) by considering the quadratic program

$$\min_{Z_k} \ \frac{1}{2}\sum_{i=1}^{J}\sum_{j=1}^{J}\left(\frac{\sum_{k\in\tau(i)} Z_k}{\bar{q}_i} - \frac{\sum_{k\in\tau(j)} Z_k}{\bar{q}_j}\right)^2 \tag{2.21}$$

$$\text{subject to} \ \sum_{k=1}^{K} \hat{M}_k Z_k = \hat{W}, \tag{2.22}$$

$$\sum_{K=1}^{K} Z_k = 1, \text{ and} \tag{2.23}$$

$$Z_k \geq 0, \text{for } k = 1, ..., K, \tag{2.24}$$

as a function of the righthand side value $\hat{W}$. We will abbreviate the objective (2.21) by $\frac{1}{2}Z^TCZ$, where C is a postive semidefinite matrix. The dual of this quadratic program (see Dorn [2]) is to choose $(Z_1, ..., Z_K, \pi_1, \pi_2)$ to

$$\max -\frac{1}{2}Z^TCZ - \hat{W}\pi_1 - \pi_2 \tag{2.25}$$

$$\text{subject to} \ \ CZ - \hat{M}\pi_1 - e^T\pi_2 \geq 0, \tag{2.26}$$

where $\hat{M} = (\hat{M}_1, ..., \hat{M}_K)^T$ and $e = (1, ..., 1)^T$. Since the objective function in (2.21) is convex and the constraints (2.22)-(2.23) are linear, it follows that there is no duality gap. Furthermore, the dual objective function (2.25), which we denote by $h(\hat{W})$, is convex with respect to $\hat{W}$ on the interval $[a^*, b^*]$. Also, $h(\hat{W})$ achieves a minimum of zero on the interval $[a_0, b_0]$. By the convexity of $h(\hat{W})$, it follows from Taksar [23] that a control limit policy is indeed optimal for (2.2)-(2.6). Thus, under the policy characterized by the interval $[a, b]$, the value of the objective function (2.6), which we denote by $f(a, b)$, will be

$$f(a,b) = \begin{cases} \frac{c\mu}{\rho_2(e^{\nu(b-a)}-1)} + \int_a^b \frac{h(x)\nu e^{\nu(x-a)}}{(e^{\nu(b-a)}-1)}dx & \text{if } \rho_1 \neq \rho_2; \\ \frac{c\sigma^2}{2(b-a)} + \int_a^b \frac{h(x)}{(b-a)}dx & \text{if } \rho_1 = \rho_2. \end{cases} \tag{2.27}$$

Our solution can then be found by minimizing $f(a, b)$ subject to $a^* \leq a \leq a^0$ and $b^0 \leq b \leq b^*$. In order to develop a sequencing policy, we propose to proceed exactly as in the limiting case $c = 0$, except use the interval $[a, b]$ derived from the solution to (2.27) in place of the interval $[a^0, b^0]$ derived from (2.12)-(2.15).

3. A Network With Controllable Inputs

The network considered in this section is identical to the closed network described in the previous section, except for the manner in which customers are released into the system. Here, customers are endogenously released according to the control process $\{N(t), t \geq 0\}$, where $N(t)$ is the number of customers released into the network up to time t. There is a constraint that the long run expected average number of customers departing the network per unit of time is at least $\bar{\lambda}$. The mix of customers released into the network is again exogenously generated according to the product mix $q = (q_k)$. As in Section 2, $v_i = \sum_{k=1}^{K} M_{ik}$ for $i = 1, 2$, but the traffic intensities are now defined by

$$\rho_i = v_i \bar{\lambda} \text{ for } i = 1, 2. \tag{3.1}$$

The balanced heavy loading conditions for the network assume the existence of a large integer n such that $\sqrt{n}(1 - \rho_i)$ is positive and of moderate size for $i = 1, 2$.

Define the scaled *input process* $\{\theta(t), t \geq 0\}$ by

$$\theta(t) = \frac{\bar{\lambda} n t - N(nt)}{\sqrt{n}}, \ t \geq 0. \tag{3.2}$$

and define the scaled queue length process $Z = (Z_k)$ and scaled idleness process $U = (U_1, U_2)$ as in Section 1. The workload formulation (see Section 3 of Wein [26]) is to choose the RCLL processes (U, Z, θ) that are nonanticipating with respect to B to

$$\min \limsup_{T \to \infty} \frac{1}{T} E\left[c \int_0^T \sum_{k=1}^{K} Z_k(t) dt + \frac{1}{2} \int_0^T \sum_{i=1}^{J} \sum_{j=1}^{J} \left(\frac{\sum_{k \in \tau(i)} Z_k(t)}{\bar{q}_i} - \frac{\sum_{k \in \tau(j)} Z_k(t)}{\bar{q}_j} \right)^2 dt \right] \tag{3.3}$$

$$\text{subject to } \sum_{k=1}^{K} M_{ik} Z_k(t) = B_i(t) + U_i(t) - v_i \theta(t) \quad \text{for all } t \geq 0 \text{ and } i = 1, 2, \tag{3.4}$$

$$\limsup_{T \to \infty} \frac{1}{T} E[U_i(T)] \leq \gamma_i \quad \text{for } i = 1, 2, \tag{3.5}$$

$$Z_k(t) \geq 0 \text{ for } k = 1, \ldots, K \text{ and for all } t \geq 0, \text{ and} \tag{3.6}$$

$$U_1 \text{ and } U_2 \text{ are nondecreasing with } U_1(0) = U_2(0) = 0, \tag{3.7}$$

where B is a two-dimensional Brownian motion and $\gamma_i = \sqrt{n}(1-\rho_i), i = 1, 2$.

As in Wein [25], this problem can be analyzed by solving for the control process Z in terms of the control process U. In particular, we can solve for Z by solving, at each time t, the following quadratic program:

$$\min \; c\sum_{k=1}^{K} Z_k(t) + \frac{1}{2}\sum_{i=1}^{J}\sum_{j=1}^{J}\left(\frac{\sum_{k\in\tau(i)} Z_k(t)}{\bar{q}_i} - \frac{\sum_{k\in\tau(j)} Z_k(t)}{\bar{q}_j}\right)^2 \tag{3.8}$$

$$\text{subject to} \; \sum_{k=1}^{K} \hat{M}_k Z_k(t) = \hat{B}(t) + \rho_2 U_1(t) - \rho_1 U_2(t) \text{ for } t \geq 0, \tag{3.9}$$

$$Z_k(t) \geq 0, \text{for } k = 1, \ldots, K. \tag{3.10}$$

where $\hat{M}_k$ is defined in (2.1), and $\hat{B}(t) = \rho_2 B_1(t) - \rho_1 B_2(t), t \geq 0$, is a one-dimensional Brownian motion with drift μ and variance σ^2. At each time t, this quadratic program has different values for the right side of (3.9), which is again denoted by $\hat{W}(t)$. The objective function will be abbreviated by $c^T Z(t) + \frac{1}{2} Z(t)^T C Z(t)$, where the matrix C is positive semidefinite.

The dual of this quadratic program is to choose $Z(t)$ and $\pi(t)$ to

$$\max \; -\frac{1}{2} Z^T C Z + \hat{W}\pi \tag{3.11}$$

$$\text{subject to} \; \hat{M}\pi - CZ \leq e^T c. \tag{3.12}$$

Once again, there is no duality gap, and the optimal dual objective function $h(\hat{W})$ is convex with respect to $\hat{W}$ and achieves a minimum of zero at zero.

Given an optimal value of $Z^*(t)$, the following constrained singular control problem is then solved: choose nondecreasing, RCLL, nonanticipating (with respect to $\hat{B}$) processes U_1 and U_2 to

$$\min \; \limsup_{T\to\infty} \frac{1}{T} E\left[\int_0^T h(\hat{W}(t))dt\right] \tag{3.13}$$

$$\text{subject to} \; \hat{W}(t) = \hat{B}(t) + \rho_2 U_1(t) - \rho_1 U_2(t) \text{ for } t \geq 0, \tag{3.14}$$

$$\limsup_{T\to\infty} \frac{1}{T} E_x[U_1(T)] = \frac{(1-\rho_1)\mu}{\rho_1 - \rho_2}, \tag{3.15}$$

$$\limsup_{T\to\infty} \frac{1}{T} E_x[U_2(T)] = \frac{(1-\rho_2)\mu}{\rho_1 - \rho_2}, \tag{3.16}$$

if $\hat{B}$ has drift $\mu \neq 0$. If $\rho_1 = \rho_2$, the $\hat{B}$ is driftless, and the right sides of (3.15) and (3.16) are replaced by $\sqrt{n}\rho_1(1-\rho_1)$.

Let us summarize the solution procedure to the workload formulation (3.3)-(3.7). The controller observes a two-dimensional Brownian motion B, from which can be observed the one-dimensional Brownian motion $\hat{B} = \rho_2 B_1 - \rho_1 B_2$. The solution (Z^*, U^*, θ^*) is given by the solution Z^* to the quadratic program (3.8)-(3.10), which depends on the process $\hat{W}$, the solution U^* to the Brownian control problem (3.13)-(3.16), and

$$\theta^*(t) = v_1^{-1}[B_1(t) + U_1^*(t) - \sum_{k=1}^{K} M_{1k} Z_k^*(t)] \text{ for all } t \geq 0. \tag{3.17}$$

We now turn to solving the constrained control problem (3.13)-(3.16). In Wein [25], a Langrangian approach was used to put the constraints (3.15)-(3.16) in the objective function (with multipliers r and l for (3.15) and (3.16), respectively) for the limiting case $c \to \infty$. By Theorem 6.2 of Wein [25], the following conditions are sufficient for optimality of (3.13)-(3.16). Let the infinitesimal generator Γ of $\hat{B}$ be given by

$$\frac{1}{2}\sigma^2 \frac{d^2}{dx^2} + \mu \frac{d}{dx}. \tag{3.18}$$

Proposition 4. *Suppose* $(g, V(x), r, l, a, b)$ *satisfy*

$$\text{Min } \{\Gamma V(x) + h(x) - g, r + V'(x), l - V'(x)\} = 0, \tag{3.19}$$

$$V(0) = 0, \tag{3.20}$$

$$\Gamma V(x) + h(x) - g = 0 \quad \text{for } a \leq x \leq b, \tag{3.21}$$

$$V'(x) = -r \text{ for } x \leq a, \tag{3.22}$$

$$V'(x) = l \quad \text{for } x \geq b, \tag{3.23}$$

$$\limsup_{T\to\infty} \frac{1}{T} E_x[U_1(T)] = \frac{(1-\rho_1)\mu}{\rho_1-\rho_2}, \quad \text{and} \tag{3.24}$$

$$\limsup_{T\to\infty} \frac{1}{T} E_x[U_2(T)] = \frac{(1-\rho_2)\mu}{\rho_1-\rho_2}, \tag{3.25}$$

where

$$U_1(t) = \frac{1}{\rho_2} \sup_{0\le s\le t} [a - \hat{B}(s) + \rho_1 U_2(s)]^+, \quad \text{and} \tag{3.26}$$

$$U_2(t) = \frac{1}{\rho_1} \sup_{0\le s\le t} [\hat{B}(s) + \rho_2 U_1^*(s) - b]^+. \tag{3.27}$$

Suppose $V \in \mathbf{C}^2$ and there exist constants N_1 and N_2 such that $V(x) \le N_1 + N_2 h(x)$. Then the optimal policy to the constrained problem (3.13)-(3.16) is (3.26)-(3.27).

Following Section 7 of Wein [25], we use Propositions 2 and 3 in Section 2 to develop a candidate solution to (3.13)-(3.16), which is characterized by the interval endpoints a^* and b^*. In particular, Proposition 2 is used to show that constraints (3.15) and (3.16) are satisfied with equality if and only if the interval width satisfies

$$b^* - a^* = \begin{cases} \nu^{-1} \ln\left(\frac{\rho_1(1-\rho_2)}{\rho_2(1-\rho_1)}\right) & \text{if } \rho_1 \neq \rho_2; \\ \frac{\sigma^2}{2\sqrt{n}\rho_1(1-\rho_1)} & \text{if } \rho_1 = \rho_2. \end{cases} \tag{3.28}$$

Proposition 3 is then used to show that the objective (3.13) is minimized by solving for a^* to minimize $f(a)$, where

$$f(a) = \begin{cases} \int_a^{a+\nu^{-1}\ln\left(\frac{\rho_1(1-\rho_2)}{\rho_2(1-\rho_1)}\right)} \frac{h(x)\nu e^{\nu(x-a)}}{(e^{\nu(b-a)}-1)} dx & \text{if } \rho_1 \neq \rho_2; \\ \frac{2\sqrt{n}\rho_1(1-\rho_1)}{\sigma^2} \int_a^{a+\frac{\sigma^2}{2\sqrt{n}\rho_1(1-\rho_1)}} h(x)dx & \text{if } \rho_1 = \rho_2. \end{cases} \tag{3.29}$$

Given the candidate policy (a^*, b^*), the obvious candidate for the gain g^* that satisfies the optimality equations is the optimal long run expected average cost function of the Lagrangian, or

$$g^* = \begin{cases} f(a^*) + r\frac{(1-\rho_1)\mu}{\rho_1-\rho_2} + \frac{(1-\rho_2)\mu}{\rho_1-\rho_2} & \text{if } \rho_1 \neq \rho_2; \\ f(a^*) + (r+l)\sqrt{n}\rho_1(1-\rho_1) & \text{if } \rho_1 = \rho_2. \end{cases} \tag{3.30}$$

The potential function $V^*(x)$ and the multipliers r^* and l^* that satisfy the optimality equations can then be found as in Section 8 of Wein [25]. A closed form solution (U_1^*, U_2^*)

to the constrained problem (3.13)-(3.16) and a proof of optimality is given in [25] for the limiting case $c \to \infty$. The solution to (3.29) is more difficult in our general case where $0 \leq c < \infty$ (a numerical example is carried out in the next section), and the verification of optimality needs to be done on a case-by-case basis.

Now let us interpret the optimal solution to the workload formulation in terms of the original queueing system. The solution $(Z^*(t), \pi^*(t))$ to the dual quadratic program (3.11)-(3.12) yields the dynamic reduced costs

$$\bar{c}_k(t) = c + (CZ^*)_k(t) - \hat{M}_k \pi^*(t) \text{ for } k = 1, ..., K, \text{ and } t \geq 0, \tag{3.31}$$

where $(CZ^*)_k(t)$ denotes the kth element of $CZ^*(t)$. This value measures the increase in the objective function of problem (3.3)-(3.7) per unit of increase in the righthand side value of the nonnegativity constraint $Z_k(t) \geq 0$. Thus, the higher the value of $\bar{c}_k(t)$, the more costly it is to hold a class k customer in queue at time t. As in Wein [26], we propose the policy that awards higher priority at time t to the classes with the higher dynamic reduced costs $\bar{c}_k(t)$. If more than one class at the same station have a dynamic reduced cost equal to zero, then we use the tie-breaking rule proposed in Yang [31] and also described in Wein [27].

The proposed input policy is to release a customer into the network whenever the two-dimensional workload process enters a specific region in the nonnegative orthant of R^2. This region is derived from the interval endpoints $[a^*, b^*]$ and from the solution Z^* of the quadratic program. The policy is most easily described with a concrete example, and so we defer its description until the next section.

4. An Example

We will illustrate the procedure described in Sections 2 and 3 with the simple example displayed in Figure 1, which was first studied in Harrison and Wein [9] and Wein [26]. There are two customer types, A and B, and type A customers have two stages on their route,

while type B customers have four stages. Thus, there are six customer classes, and they will be designated (and ordered from $k = 1, ..., 6$) by $A1, A2, B1, B2, B3$, and $B4$. For concreteness, all processing times are assumed to be exponential, and the mean processing times for all six classes are displayed in Figure 1. The specified product mix is $q = (1/2, 0, 1/2, 0, 0, 0)$, and customers are released into the network in the order $ABABAB...$

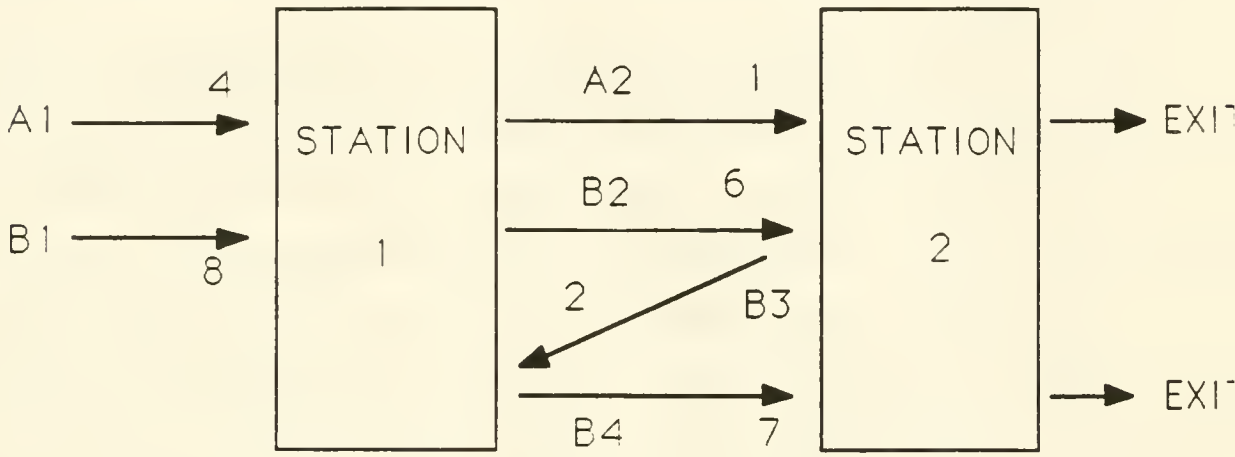

Figure 1. An example.

The workload profile matrix $M = (M_{ik})$ is given by

$$M = \begin{pmatrix} 4 & 0 & 10 & 2 & 2 & 0 \\ 1 & 1 & 13 & 13 & 7 & 7 \end{pmatrix}. \tag{4.1}$$

where M_{ik} is the expected remaining processing time at station i for a class k customer until that customer exits. Also, the drift μ and variance σ^2 of the one-dimensional Brownian motion $\hat{B}$ was calculated in Wein [26] to be zero and 10.93, respectively.

For the closed network problem of Section 2, we have $v_1 = v_2 = 7$ so that $\rho_1 = \rho_2 = 1$. Therefore, when $c \to \infty$, the proposed sequencing policy ranks each customer class by the index

$$\hat{M}_k = (3 \ -1 \ -3 \ -11 \ -5 \ -7) \quad \text{for } k = 1, ..., 6, \tag{4.2}$$

and gives priority (from highest to lowest) in the order (B3, B1, A1) at station 1 and (A2, B4, B2) at station 2.

Turning now to the limiting case $c = 0$, it can be calculated that the interval $[a^0, b^0]$ satisfying (2.12)-(2.15) is $[-6, 0]$, which is strictly contained in the interval $[a^*, b^*] = [-11, 3]$, which is calculated from (4.2). When $\hat{W} = -6$, the solution is $Z^{a^0} = (0, 1/2, 0, 1/2, 0, 0)$ and when $\hat{W} = 0$, the solution is $Z^{b^0} = (1/2, 0, 1/2, 0, 0, 0)$. Thus we give top priority to class $B3$ at station 1 and class $B4$ at station 2, since these classes are in the higher priority bracket.

In order to decide how to award priorities among the classes that are in the lower priority bracket at each station, a simple dynamic scheme is used that attempts to keep the number of customers in queue of each class in the relative proportions dictated by Z^{a^0} and Z^{b^0}. For simplicity, we will describe this heuristic in the case where there are two classses, call them classes 1 and 2, in the lower priority bracket of a given station. Let Z_1 and Z_2 be the positive values of these classes from the solution Z^{a^0} and Z^{b^0}. The idea behind the heuristic is to keep the proportion of class 1 customers in queue to be $Z_1/(Z_1 + Z_2)$. Suppose at time t, there are $Q_1(t) > 0$ class 1 customers and $Q_2(t) > 0$ class 2 customers in queue, and the server has just completed a service and is ready to begin serving a new customer. If the server serves a class 1 customer with probability $p(t)$ and a class 2 customer with probability $1 - p(t)$, then, after the choice is made, the expected proportion of total jobs in queue that are of class 1 is

$$\frac{p(t)(Q_1(t) - 1) + (1 - p(t))Q_1(t)}{Q_1(t) + Q_2(t) - 1}. \tag{4.3}$$

Setting this equal to the desired proportion $Z_1/(Z_1 + Z_2)$ and solving for $p(t)$ yields

$$p(t) = \frac{Z_1 - Z_2 Q_1(t) - Z_1 Q_2(t)}{Z_1 + Z_2}. \tag{4.4}$$

The dynamic heuristic is to serve class 1 customers at time t if $p(t) > 1$, serve class 2 customers if $p(t) < 0$, and serve class 1 customer with probability $p(t)$ if $p(t) \in (0, 1)$.

Returning to our example, from the solution Z^{a^0} (respectively, Z^{b^0}), it is desirable to keep the queue lengths of classes $A2$ and $B2$ (respectively, $A1$ and $B1$) at station 2

(respectively, station 1) as equal as possible, and thus our dynamic heuristic described in (4.4) takes on a particularly simple form. When no class $B3$ customers are available for processing at station 1, we serve class $A1$ customers when $Q_1(t) > Q_3(t)$ and serve class $B1$ customers when $Q_3(t) > Q_1(t)$, and alternate (or flip a fair coin) when there is a tie. Similarly, when there are no class $B4$ customers in queue at station 2, we serve class $A2$ customers when $Q_2(t) > Q_4(t)$ and serve $B2$ class customers when $Q_4(t) > Q_2(t)$.

For the general case where $0 < c < \infty$, the optimal dual objective function $h(\hat{W})$ in (2.25) is given by

$$h(\hat{W}) = \begin{cases} (\frac{6+\hat{W}}{5})^2 & \text{if } \hat{W} \in [-11, -6]; \\ 0 & \text{if } \hat{W} \in [-6, 0]; \\ \frac{\hat{W}^2}{9} & \text{if } \hat{W} \in [0, 3], \end{cases} \tag{4.5}$$

and thus $f(a, b)$ is given by

$$f(a, b) = \frac{1}{b-a}\left(\frac{c\sigma^2}{2} + \frac{b^3}{27} - \frac{a^3}{75} - \frac{18a^2}{75} - \frac{108a}{75} - \frac{216}{75}\right). \tag{4.6}$$

Therefore, our optimal interval endpoints can be found by minimizing $f(a, b)$ over $a \in [-11, -6]$ and $b \in [0, 3]$.

Now let us consider the numerical example in the context of a network with controllable inputs. If we choose a long run average throughput rate of .1286 customers per unit of time and choose the scaling parameter $n = 100$, then $\rho_1 = \rho_2 = .9$. The solution $(Z^*(t), \pi^*(t))$ to the dual quadratic program (3.11)-(3.12) is given in Table I, and the dynamic reduced costs (3.31) are given in Table II. In each of these tables, the solution and costs have been broken down into three regions (corresponding to the columns in the two tables), depending upon the value of the workload imbalance process $\hat{W}(t)$. The optimal dual objective function $h(\hat{W})$ in (3.11) is given by

$$h(\hat{W}) = \begin{cases} -\frac{25c^2}{144} - \frac{c\hat{W}}{6} & \text{if } \hat{W} \leq -55c/12; \\ -\frac{c\hat{W}}{11} + \frac{\hat{W}^2}{121} & \text{if } \hat{W} \in [-55c/12, 0]; \\ \frac{c\hat{W}}{3} + \frac{\hat{W}^2}{9} & \text{if } \hat{W} \geq 0. \end{cases} \tag{4.7}$$

By (3.28), it follows that $b^* - a^* = 6.072$, which we will denote by k. Solving (3.29) yields

$$a^* = \begin{cases} -\frac{9}{4}c - k + \sqrt{\frac{7}{2}c^2 + \frac{3}{2}ck} & \text{if } c \in [0, 6k/35]; \\ -\frac{11}{14}k & \text{if } c \geq 6k/35. \end{cases} \tag{4.8}$$

VARIABLE	$\hat{W}(t) \leq -55c/12$	$\hat{W}(t) \in [-55c/12, 0]$	$\hat{W}(t) \geq 0$
$Z_1^*(t)$	0	0	$\frac{\hat{W}(t)}{3}$
$Z_2^*(t)$	$-\frac{\hat{W}(t)}{12} - \frac{55c}{144}$	0	0
$Z_3^*(t)$	0	0	0
$Z_4^*(t)$	$-\frac{\hat{W}(t)}{12} + \frac{5c}{144}$	$-\frac{\hat{W}(t)}{11}$	0
$Z_5^*(t)$	0	0	0
$Z_6^*(t)$	0	0	0
$\pi^*(t)$	$-\frac{c}{6}$	$\frac{2\hat{W}(t)}{121} - \frac{c}{11}$	$\frac{c}{3} + \frac{2\hat{W}(t)}{9}$

TABLE I. Solution to the Dual Quadratic Program (3.11)-(3.12).

In order to interpret the solution to the workload formulation given in Table I and (4.8), let us consider the limiting example $c \to \infty$ analyzed in Wein [26]; readers may verify that the solution derived here in the limiting case is identical to the solution found in Section 7 of Wein [26]. In this case, there are only two regions (i.e., two columns) to consider in Tables I and II, and it is easily seen that the sequencing policy based on the dynamic reduced costs in Table II ranks all classes by the index $\hat{M}_k$ in (4.2), and serves the class with the smallest (respectively, largest) value of the index when $\hat{W}(t) > 0$ (respectively, $\hat{W}(t) < 0$). From Table 1, we see that only one component of Z^* is positive at any time t and only two components are ever positive. Thus the two-dimensional workload process defined by $W_i(t) = \sum_{k=1}^{6} M_{ik} Z_k(t)$ for $i = 1, 2$ and $t \geq 0$, stays on the boundary of a cone in the nonnegative orthant of R^2. Furthermore, the interval $[a^*, b^*]$, which equals $[-4.771, 1.301]$ by (4.8), determines cutoff points on the cone boundary beyond which the workload process may not enter (see Figure 2). As explained in Wein [26], the control process $\theta(t)$, which can move either way along the 45 degree direction in Figure 2 (see (3.4)), is used to keep the workload process on the truncated cone boundary in Figure 2.

Thus, when the workload process is in the region to the lower left of the truncated cone, then exerting ϵ corresponds to releasing more jobs into the system relative to the nominal input rate $\bar{\lambda}$.

CLASS	$\dot{\hat{W}}(t) \leq -55c/12$	$\dot{\hat{W}}(t) \in [-55c/12.0]$	$\dot{\hat{W}}(t) \geq 0$
$A1$	$\frac{2c}{3}$	$\frac{14c}{11} + \frac{16\dot{\hat{W}}(t)}{121}$	0
$A2$	0	$\frac{10c}{11} + \frac{24\dot{\hat{W}}(t)}{121}$	$\frac{4c}{3} + \frac{8\dot{\hat{W}}(t)}{9}$
$B1$	$\frac{4c}{3}$	$\frac{8c}{11} - \frac{16\dot{\hat{W}}(t)}{121}$	$2c$
$B2$	0	0	$\frac{14c}{3} + \frac{16\dot{\hat{W}}(t)}{9}$
$B3$	c	$\frac{6c}{11} - \frac{12\dot{\hat{W}}(t)}{121}$	$\frac{8c}{3} + \frac{4\dot{\hat{W}}(t)}{9}$
$B4$	$\frac{2c}{3}$	$\frac{4c}{11} - \frac{8\dot{\hat{W}}(t)}{121}$	$\frac{10c}{3} + \frac{8\dot{\hat{W}}(t)}{9}$

TABLE II. Dynamic Reduced Costs.

The resulting release policy is called a *workload regulating* release policy, and it releases a job into the network whenever the unscaled workload process $w(t)$, where $\hat{W}(t) = w(nt)/\sqrt{n}$, enters a specific region. For our example, this region, which is shaded in Figure 2, is

$$w_1(t) \leq 19 \quad \text{and} \tag{4.9}$$

$$w_2(t) - \frac{1}{4}w_1(t) \leq \frac{3}{4}\epsilon, \tag{4.10}$$

or

$$w_2(t) \leq 62 \quad \text{and} \tag{4.11}$$

$$w_1(t) - \frac{2}{13}w_2(t) \leq \frac{11}{13}\epsilon, \tag{4.12}$$

where the parameter ϵ, which shifts the cone vertex from the origin to the point (ϵ, ϵ), is chosen so that the desired throughput rate is met. Choosing $\epsilon = 1$ achieved the target

throughput rate in the simulation run that appears later in this section. A more detailed explanation that interprets the solution (Z^*, U^*, θ^*) of the workload formulation (3.3)-(3.7) in order to obtain the workload regulating release policy defined by (4.9)-(4.12) can be found in Section 6 of Wein [26].

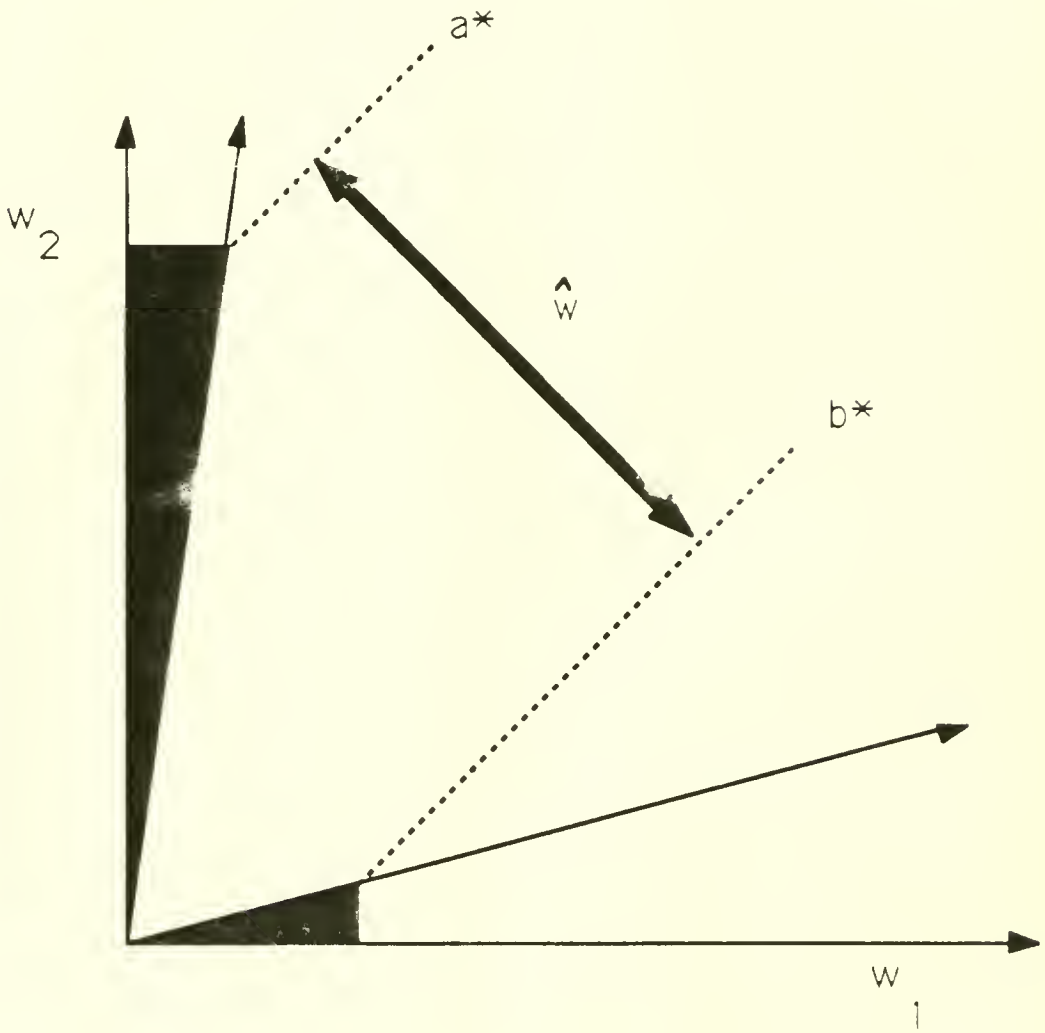

Figure 2. The release region when $c \to \infty$.

For the case when $0 \leq c < \infty$, the sequencing policy can easily be calculated from the dynamic reduced costs in Table II, but the determination of the release policy is more complicated. In the limiting case $c = 0$, equation (4.8) implies that $[a^*, b^*] = [-6.072, 0]$

and thus $\hat{W}(t)$ will always be in the leftmost region in Table I. From this table, it follows that $Z_2^*(t) = Z_4^*(t) = -\hat{W}(t)/12$, and so the workload process stays on a segment of the line $7W_1(t) = W_2(t)$. The release policy in this case (see Wein [26] for an explanation) is to release a job whenever

$$w_2(t) \leq 71 \quad \text{and} \quad 7w_1(t) - w_2(t) \leq 6\epsilon, \tag{4.13}$$

which is pictured in Figure 3. Once again, $\epsilon = 1$ achieved the desired throughput rate in the simulation experiment.

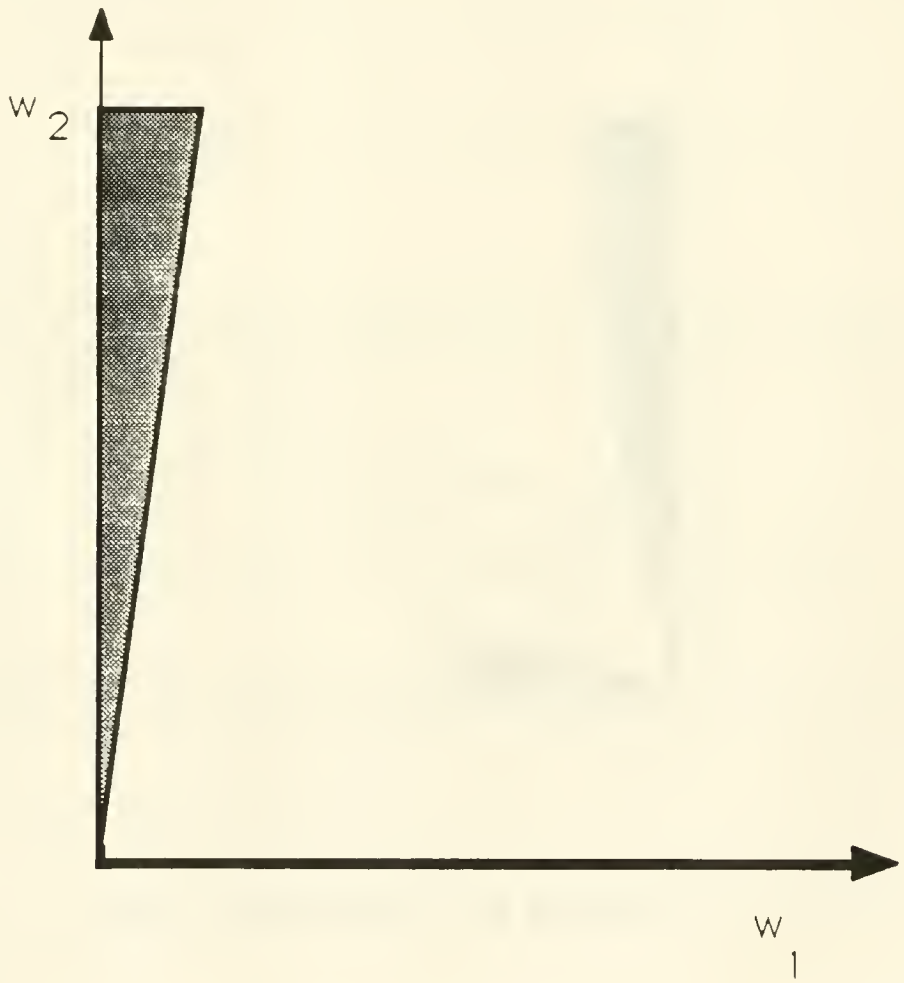

Figure 3. The release region when $c = 0$.

When $0 < c < \infty$, then the release region will be as pictured in Figure 4. The slope

of the lower ray is 1/4, just as it is in Figure 2. The upper shaded region in Figure 4 is formed by a line that has a slope of 13/2 (the same as the upper ray in Figure 2) and another line that has a slope of 7 (which is parallel to the upper ray in Figure 3.) The intersection of these two lines is at $\hat{W}(t) = -55c/12$, which is the cutoff point between two of the regions in Table I.

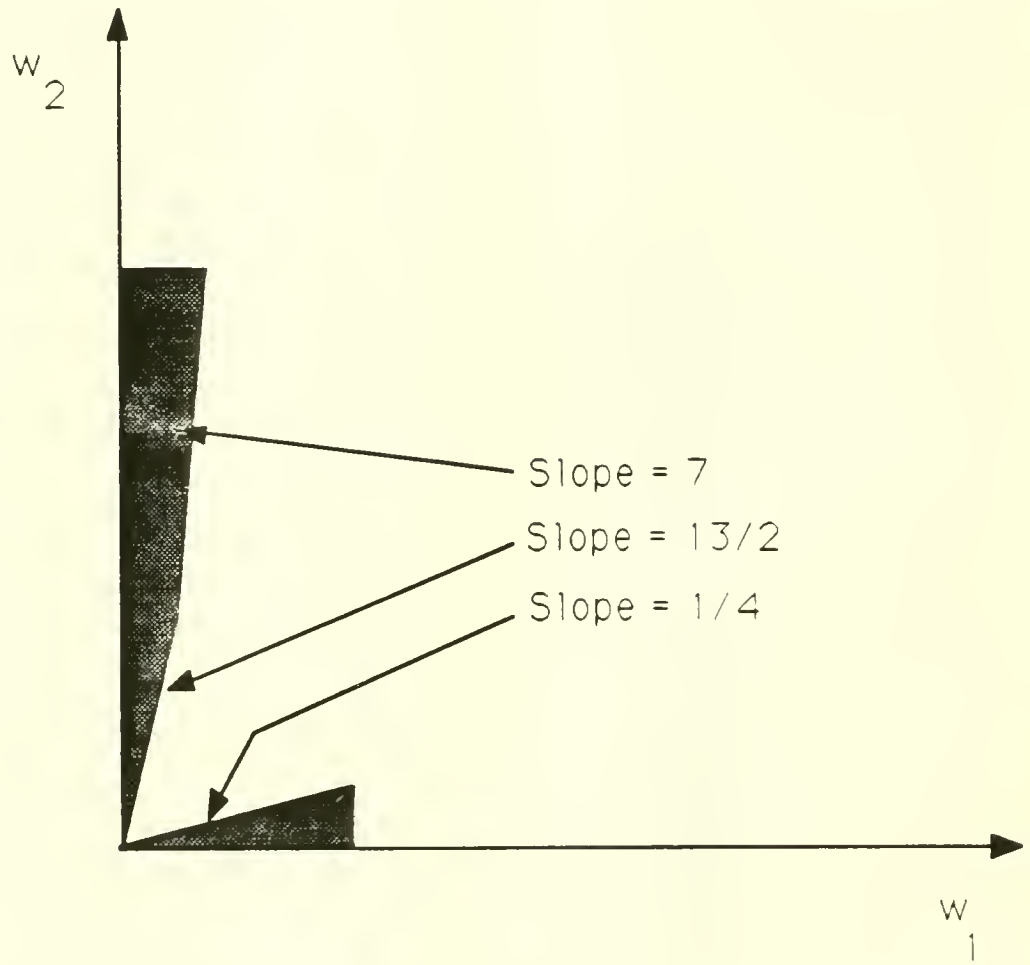

Figure 4. The release region when $0 < c < \infty$.

Before turning to the simulation results, there is one refinement that has been made to the scheduling policies described earlier. Notice in Table II that many of the scaled queue length processes are zero. Due to the rescaling, Z measures how many *tens* of

customers are in the actual network. This measurement is obviously too crude to account for the customers who are in service, and so we propose a slight refinement to the surrogate sojourn time variance that essentially reinterprets Z as only the scaled number of customers in queue, not including the customers in service. This refinement will hopefully cause a reduction in the surrogate sojourn time variance, and it will be described in terms of the numerical example. For this numerical example, each server is to be utilized about 90% of the time, and we will denote this by $\rho = .9$. From Figure 1, the mean number of type A customers in service is $5\rho/14$ and the mean number of type B customers in service is $23\rho/14$. The refinement changes the surrogate sojourn variance from

$$\left(\frac{Z_1(t) + Z_2(t)}{1/2} - \frac{Z_3(t) + Z_4(t) + Z_5(t) + Z_6(t)}{1/2} \right)^2 \tag{4.14}$$

in (2.21) and (3.8) to

$$\left(\frac{Z_1(t) + Z_2(t) + \frac{5\rho}{14N}}{1/2} - \frac{Z_3(t) + Z_4(t) + Z_5(t) + Z_6(t) + \frac{23\rho}{14N}}{1/2} \right)^2 \tag{4.15}$$

in (2.21) and

$$\left(\frac{Z_1(t) + Z_2(t) + \frac{5\rho}{14\sqrt{n}}}{1/2} - \frac{Z_3(t) + Z_4(t) + Z_5(t) + Z_6(t) + \frac{23\rho}{14\sqrt{n}}}{1/2} \right)^2 \tag{4.16}$$

in (3.8). In our simulation experiment, this refinement was only analyzed for the case $c = 0$ in the closed network. This is a particularly easy case to evaluate, since the only difference in the analysis is to change equation (2.15) from

$$Z_1 + Z_2 - Z_3 - Z_4 - Z_5 - Z_6 = 0 \tag{4.17}$$

to

$$Z_1 + Z_2 - Z_3 - Z_4 - Z_5 - Z_6 = \frac{18\rho}{14N}. \tag{4.18}$$

A simulation experiment was performed to assess the effectiveness of the analysis presented in Sections 2 and 3. Two conventional policies, which consist of a customer

release policy and a priority sequencing policy, were tested. One policy was deterministic input (abbreviated by DET in Table III), where customers are released at constant intervals in the order $ABABAB...$, paired with FCFS sequencing, and the second was closed loop input (CL), where the total population level was held constant, and FCFS. The closed release policy was also tested in conjunction with the sequencing policies derived in Section 2 under the two limiting cases, $c \to \infty$ and $c = 0$. The $c = 0$ policy used the refinement described in equation (4.18). The job release and priority sequencing policies derived in Section 3, which will be denoted by WR (for workload regulating), were also tested under the two limiting cases.

In order to allow for a comparison of the various cases, we have set the parameters N and ϵ so that each policy achieves the same average throughput rate, which for convenience was chosen to be .127 customers per unit of time, which corresponds to a server utilization of 88.9%. Due to the discrete nature of the population parameter N, we were not always able to obtain this target rate exactly. In such cases, we have reported linear interpolations of the various performance measures so that the average throughput rate would be .127. For each policy tested, 20 independent runs were made, each consisting of 5000 customer completions and no initialization period. The mean sojourn time, the *actual* sojourn time standard deviation (as opposed to the surrogate sojourn time standard deviation), and the mean throughput rate, along with 95% confidence intervals are recorded in Table III.

Referring to the results in Table III, it is seen that the (DET,FCFS) policy achieved a sojourn time standard deviation that was twice as high as most of the other policies, implying that job release has a large impact on the sojourn time variance. Of course, in many manufacturing systems, the amount of time a job spends waiting to gain entrance onto the shop floor is also of importance, and that time is ignored in this study; readers are referred to Wein and Chevalier [28] for a scheduling study where this time is taken into account. Minimizing the surrogate sojourn time variance appears to be an effective means of reducing the actual sojourn time variance, since the two $c = 0$ cases reduced the sojourn

time variance relative to the corresponding $c \to \infty$ cases. The (CL.$c = 0$) case achieved the lowest sojourn time variance, perhaps in part because of the refinement described in (4.18).

SCHEDULING POLICY	MEAN SOJOURN TIME	VARIANCE OF SOJOURN TIME	MEAN THROUGHPUT
DET,FCFS	90.6 (±5.0)	61.3 (±3.83)	.127 (±.0000)
CL,FCFS	71.7 (±.45)	31.6 (±.54)	.127 (±.0008)
CL,$c \to \infty$	50.4 (±.31)	27.9 (±.27)	.127 (±.0009)
CL,$c = 0$	68.9 (±.43)	25.4 (±.26)	.127(±.0008)
WR,$c \to \infty$	39.1 (±.52)	33.0 (±.53)	.127 (±.0009)
WR,$c = 0$	53.3 (±1.45)	29.5 (±.71)	.127 (±.0008)

TABLE III. Simulation Results for Example 1.

For this particular example, the difference in mean sojourn times between the $c = 0$ and $c \to \infty$ cases is much larger than the corresponding difference in sojourn time variance. Thus, relative to the $c \to \infty$ case, the queueing system would have to incur a large increase in mean sojourn time in order to achieve a relatively small decrease in sojourn time standard deviation. However, this phenomena is problem specific, and can be explained by analyzing the sequencing policy for the (CL,$c \to \infty$) case given in (4.2), and by recalling the following fact about priority queueing systems. When a queueing system is heavy loaded and static priorities are used, then the bottom priority class at each server incurs much more delay in queue than the other classes; as mentioned earlier, this has been quantified in several heavy traffic limit theorems. Since product A gets bottom priority at station 1 and product B gets bottom priority at station 2 under the (CL,$c \to \infty$) case, the surrogate sojourn time variance for this policy was not that large, and was less than it was under the (CL,FCFS) case. Hence, for this particular example, the (CL,$c \to \infty$) case provides

a simultaneous reduction in the mean and variance of the sojourn times. Furthermore, a large reduction in sojourn time variance relative to the (CL,$c \rightarrow \infty$) case is not possible.

On the other hand, consider our same example, but now change the mean processing times for the six classes from (4, 1, 8, 6, 2, 7) to (2, 6, 4, 1, 8, 7); we will denote the example using this new data set as example 2. Now the (CL,$c \rightarrow \infty$) policy awards lowest priority to type B products at both stations, and so the surrogate variance is now very large for this policy, and is larger than the (CL,FCFS) case. Therefore, we might expect that the difference in sojourn time variances between the two cases (CL,$c \rightarrow \infty$) and (CL,$c = 0$) will be larger than for our original example.

SCHEDULING POLICY	MEAN SOJOURN TIME	VARIANCE OF SOJOURN TIME	MEAN THROUGHPUT
CL,FCFS	71.6 (±.36)	31.6 (±.22)	.127 (±.0007)
CL,$c \rightarrow \infty$	52.8 (±.21)	37.9 (±.21)	.127 (±.0007)
CL,$c = 0$	63.8 (±.24)	23.2 (±.20)	.127 (±.0007)

TABLE IV. Simulation Results for Example 2.

Simulation results for example 2 are displayed in Table IV. As expected, the sojourn time variance for the (CL,$c \rightarrow \infty$) case is larger than the (CL,FCFS) case, and there is a large diffeence in sojourn time variance between the two extreme cases of (CL,$c \rightarrow \infty$) and (CL,$c = 0$). Thus, relative to the (CL,$c \rightarrow \infty$) case, a large decrease in sojourn time variance can be achieved while incurring a somewhat moderate increase in mean sojourn time. Also, as in the original example, example 2 offers a policy, the (CL,$c = 0$) policy, that simultaneously reduces the mean and variance of processing times relative to FCFS.

Perhaps the most interesting result from this study is that FCFS is not particularly effective at minimizing sojourn time variance in complex queueing systems. This paper offers some aid in developing release and sequencing policies to reduce the sojourn time

variance in such systems.

REFERENCES

[1] Avriel, M. (1976). *Nonlinear Programming: Analysis and Methods.* Prentice-Hall, Englewood Cliffs, New Jersey.

[2] Dorn, W. S. (1960). Duality in Quadratic Programming. *Q. Appl. Math.* 18, 155-162.

[3] Federgruen, A. and Grownevelt, H. (1988). $M/G/c$ Queueing Systems with Multiple Customer Classes: Characterization and Control of Achievable Performance under Nonpreemptive Priority Rules. *Management Science* 9, 1121-1138.

[4] Groenevelt, H. (1989). Private Correspondence.

[5] Harrison, J. M. (1973). A Limit Theorem for Priority Queues in Heavy Traffic. *J. Appl. Prob.* 10, 907-912.

[6] Harrison, J. M. (1985). *Brownian Motion and Stochastic Flow Systems.* John Wiley and Sons, New York.

[7] Harrison, J. M. (1988). Brownian Models of Queueing Networks with Heterogeneous Customer Populations. In W. Fleming and P. L. Lions (eds.), *Stochastic Differential Systems, Stochastic Control Theory and Applications*, IMA Volume 10, Springer-Verlag, New York, 147-186.

[8] Harrison, J. M. and Taksar, M. I. (1983). Instantaneous Control of Brownian Motion. *Math. of Oper. Res.* 3, 439-453.

[9] Harrison, J. M. and Wein, L. M. (1989). Scheduling Networks of Queues: Heavy Traffic Analysis of a Two-Station Closed Network. To appear in *Operations Research.*

[10] Jackson, J. R. (1960). Some Problems in Queueing with Dynamic Priorities. *Naval Res. Log. Quart.* 7, 235-249.

[11] Jackson, J. R. (1961). Queues with Dynamic Priorities. *Management Science* 1, 18-34.

[12] Jackson, J. R. (1960). Waiting Time Distributions for Queues with Dynamic Priorities. *Naval Res. Log. Quart.* 9, 31-36.

[13] Johnson, D. P. (1983). Diffusion Approximations for Optimal Filtering of Jump Processes and for Queueing Networks. Unpublished Ph.D. thesis, Dept. of Mathematics, Univ. of Wisconsin, Madison.

[14] Kelly, F. P. (1979). *Reversibility and Stochastic Networks*, John Wiley and Sons, New York.

[15] Kingman, J. F. C. (1962). The Effect of Queue Discipline on Waiting Time Variance. *Proc. Camb. Phil. Soc.* 58, 163-164.

[16] Kleinrock, L. (1964). A Delay Dependent Queue Discipline. *Naval Res. Log. Quart.* 11, 329-341.

[17] Kleinrock, L. (1976). *Queueing Systems Vol. II: Computer Applications*. John Wiley and Sons, New York.

[18] Klimov, G. P. (1974). Time Sharing Service Systems I. *Th. Prob. Appl.* 19,532-551.

[19] Little, J. D. C. (1961). A Proof of the Queueing Formula $L = \lambda W$. *Operations Research* 9, 383-387.

[20] Peterson, W. P. (1989). A Heavy Traffic Limit Theorem for Networks of Queues with Multiple Customer Types. To appear in *Math. Operations Research*.

[21] Reiman, M. I. (1983). Some Diffusion Approximations with State Space Collapse. *Proc. Intl. Seminar on Modeling and Performance Evaluation Methodology*, Springer-Verlag, Berlin.

[22] Shanthikumar, J. G. (1982). On Reducing Time Spent in $M/G/1$ Systems. *European J. Operational Research* 9, 286-294.

[23] Taksar, M. I. (1985). Average Optimal Singular Control and a Related Stopping Problem. *Math. Operations Research* 10, 63-81.

[24] Walrand, J. (1988). *An Introduction to Queueing Networks*. Prentice-Hall, Englewood Cliffs, New Jersey.

[25] Wein, L. M. (1989). Optimal Control of a Two-Station Brownian Network. To appear in *Math. of Operations Research*

[26] Wein, L. M. (1989). Scheduling Networks of Queues: Heavy Traffic Analysis of a Two-Station Network with Controllable Inputs. To appear in *Operations Research*.

[27] Wein, L. M. (1989). Scheduling Networks of Queues: Heavy Traffic Analysis of a Multistation Network with Controllable Inputs. Submitted to *Operations Research*.

[28] Wein, L. M. and Chevalier, P. B. (1989). A Broader View of the Job-Shop Scheduling Problem. Submitted to *Management Science*.

[29] Whitt, W. (1971). Weak Convergence Theorems for Priority Queues: Preemptive-Resume Discipline. *J. Appl. Prob.* 8, 74-94.

[30] Wood, D. and Sargent, R. (1984). The Synthesis of Multiclass Single Server Queueing Systems. Unpublished manuscript.

[31] P. Yang (1988). Pathwise Solutions for a Class of Linear Stochastic Systems. Unpublished Ph. D. Thesis, Dept. of Operations Research, Stanford University, Stanford, CA.